OLLSCOIL NA GAILLIMHE
UNIVERSITY OF GALWAY

Autumn Examinations 2022-2023

Course Instance Code(s)	4BCT1, 4BDS1, 4BMS2, 4BS2, 1MECE1, 1MEME1, SPE
Exam(s)	Fourth BSc in Computer Science & Information Technology Fourth Bachelor of Arts with Data Science Fourth Bachelor of Science (Mathematical Science) Fourth Bachelor of Science (Hons.) ME in Electronic & Computer Engineering ME in Mechanical Engineering Structured PhD
Module Code(s)	CT4101
Module(s)	Machine Learning
Paper No.	1
External Examiner(s)	Dr. Ramona Trestian
Internal Examiner(s)	Professor Michael Madden *Dr. Frank Glavin

Instructions: Answer any 3 questions out of 4 questions.
All questions carry equal marks (25 marks each).
The total (out of 75 marks) will be converted to a percentage after marking.

.

Duration	2 hours
No. of Pages	5
Discipline(s)	School of Computer Science
Course Co-ordinator(s)	Dr. Colm O'Riordan (BCT), Dr. Nick Tosh (BDS), Prof. Dane Flannery (BMS), Dr. Emma Holahan (BS), Prof. Martin Glavin (MECE), Dr. Noel Harrison (MEME)

Requirements:

Release in Exam Venue	Yes [ ]	No [X]
MCQ Answersheet	Yes [ ]	No [X]
Handout	None	
Statistical/ Log Tables	None	
Cambridge Tables	None	
Graph Paper	None	
Log Graph Paper	None	
Other Materials	None	
Graphic material in colour	Yes [ ]	No [X]

PTO

Question 1 (25 marks)

property_id	num_beds	num_baths	floor_area	sale_price
275	1	1	40	123000
314	2	1	50	150000
2212	3	4	130	265000
3390	3	2	90	205000

The training dataset above contains data about the prices achieved at a recent auction for various properties. You are required to develop a *k*-nearest neighbours (kNN) model using this training data; this model will be used to predict the target variable sale_price for properties to be sold at future auctions.

When answering each part below, you should provide detailed comments explaining your calculations.

i. In preparation for applying 0-1 normalisation, choose appropriate minimum and maximum values for each of the independent variables (num_beds, num_baths, floor_area) that will be used to make predictions.

[3]

ii. Using the minimum and maximum values that you chose for part i. above, compute the 0-1 normalised values for all independent variables for each data point in the training set. You should present the normalised data in tabular format.

[4]

iii. Choose an appropriate distance metric (or similarity index) to use when applying k-NN to this dataset, and briefly justify your choice.

[2]

iv. Using your chosen distance metric (or similarity index) from part iii. above, compute the distance (or similarity) between each pair of points in the training dataset using the normalised data from part ii.

[5]

v. Using a 3-NN model with uniform weighting and your chosen distance metric (or similarity index), compute the predicted sale_price for each instance in the training dataset.

[6]

vi. Using the predicted sale_price values from part v. above, compute the full form of the RMSE on the training set for the 3-NN model that you have developed.

[3]

vii. Briefly comment on the RMSE value achieved by your 3-NN model – does this RMSE value indicate that the model performs well on the training data? Briefly describe one modification that could be made to the algorithm to improve the RMSE.

[2]

PTO

Question 2 (25 marks)

Part (a)

Explain what is meant by the term **hyperparameter** in the context of Machine Learning algorithms. Briefly discuss how a grid search could be used to determine suitable hyperparameter values for a machine learning algorithm.

[4]

Part (b)

ID	Target	Prediction
1	Negative	Positive
2	Negative	Negative
3	Positive	Positive
4	Positive	Positive
5	Positive	Negative
6	Negative	Negative
7	Negative	Negative
8	Negative	Positive
9	Positive	Positive
10	Positive	Negative

The table above presents the results of evaluating a classifier on a test set for a binary classification task.

i. Present the results of the evaluation above in a confusion matrix [3]
ii. Calculate the misclassification rate of the classifier [2]
iii. Calculate the true negative rate of the classifier [2]
iv. Calculate the precision of the classifier [2]
v. Calculate the recall of the classifier [2]

[11]

Part (c)

Explain what is meant by the term **cross validation**. Briefly discuss how you would use cross validation to evaluate the performance of a classification task. What is meant by the **stratification** of examples and why is it important?

[5]

Part (d)

Explain how Receiver Operating Characteristic (ROC) curves may be used to compare the performance of different classifiers for a binary classification task. Sketch an example of an ROC curve for a binary classification task as part of your answer, clearly labelling the axes and clearly indicating a point that represents ideal performance.

[5]

PTO

Question 3 (25 marks)

Part (a)

An aeronautical engineer has sent you the following email message. Prepare a detailed reply in your own words.

"I am trying to predict the strength for some new composite materials in which we embed fibres in resin, based on some quantities that we can control such as the percentage by volume of fibres, the orientation of fibres, how much cooling we apply, and the overall mass of material. Since this is a numerical quantity, I believe I could use a regression algorithm for this, is that right? From some initial reading, I have heard of linear regression, polynomial regression, and logistic regression, but I am not clear about the differences between them. Can you explain the main distinctions, and let me know what kind of data I need for them? Also, can you recommend at least two other algorithms I should look into, in addition to those I have just mentioned, and explain why you would recommend them?"

[10]

Part (b)

The engineer has a dataset with 1000 cases, and initially intended to train a model with all of this data, and then test the model on all of the data also. In your own words, explain why this would not be a good approach. Continuing from this, explain what the distinctions are between data used for training and testing, and how they should divide the data. Is there a third category they should consider? If so, explain what it is and how they should handle it.

[8]

Part (c)

The engineer asks how many cases they need to fully train a model. Provide a clear response in your own words, making reference to learning curves.

[4]

Part (d)

The engineer has built two models, A and B, and wishes to determine which is best. Describe in your own words how to perform a suitable statistical test for this.

[3]

PTO

Question 4 (25 marks)

Part (a)

Explain how the **McCulloch and Pitts artificial neuron** works. You should include a labelled diagram of the neuron and all equations necessary to calculate the output of the neuron.

[6]

Part (b)

ID	CHEST PAIN	BLOOD PRESSURE	HEART DISEASE
1	FALSE	HIGH	FALSE
2	TRUE	LOW	TRUE
3	FALSE	LOW	FALSE
4	TRUE	HIGH	TRUE
5	FALSE	HIGH	FALSE

The training dataset above contains data about the incidence of heart disease in 5 different patients. ID is a unique identifier for each data point, CHEST PAIN and BLOOD PRESSURE are *independent* attributes, and HEART DISEASE is the *target* attribute.

Calculate the information gain for the attribute CHEST PAIN and the attribute BLOOD PRESSURE. How would this information be used when constructing a Decision Tree?

[9]

Part (c)

In the context of data pre-processing, explain what is meant by **binning.** Your answer should outline the difference between **equal frequency binning** and **equal width binning**. Include diagrams to aid your explanation.

[5]

Part (d)

In the context of datasets, describe the terms **bias** and **variance**. Outline the effect of having high bias versus high variance in terms of overfitting/underfitting.

[5]

END